Journal of Machine Learning Research 14 (2013) 2487-2517 Submitted 4/12; Revised 3/13; Published 9/13

Greedy Feature Selection for Subspace Clustering

Eva L. Dyer E.DYER@RICE.EDU
Department of Electrical & Computer Engineering
Rice University
Houston, TX, 77005, USA

Aswin C. Sankaranarayanan SASWIN@ECE.CMU.EDU
Department of Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Richard G. Baraniuk RICHB@RICE.EDU
Department of Electrical & Computer Engineering
Rice University
Houston, TX, 77005, USA

Editor: Tong Zhang

Abstract

Unions of subspaces provide a powerful generalization of single subspace models for collections of high-dimensional data; however, learning multiple subspaces from data is challenging due to the fact that segmentation—the identification of points that live in the same subspace—and subspace estimation must be performed simultaneously. Recently, sparse recovery methods were shown to provide a provable and robust strategy for *exact feature selection* (EFS)—recovering subsets of points from the ensemble that live in the same subspace. In parallel with recent studies of EFS with ℓ_1-minimization, in this paper, we develop sufficient conditions for EFS with a greedy method for sparse signal recovery known as orthogonal matching pursuit (OMP). Following our analysis, we provide an empirical study of feature selection strategies for signals living on unions of subspaces and characterize the gap between sparse recovery methods and nearest neighbor (NN)-based approaches. In particular, we demonstrate that sparse recovery methods provide significant advantages over NN methods and that the gap between the two approaches is particularly pronounced when the sampling of subspaces in the data set is sparse. Our results suggest that OMP may be employed to reliably recover exact feature sets in a number of regimes where NN approaches fail to reveal the subspace membership of points in the ensemble.

Keywords: subspace clustering, unions of subspaces, hybrid linear models, sparse approximation, structured sparsity, nearest neighbors, low-rank approximation

1. Introduction

With the emergence of novel sensing systems capable of acquiring data at scales ranging from the nano to the peta, modern sensor and imaging data are becoming increasingly high-dimensional and heterogeneous. To cope with this explosion of high-dimensional data, one must exploit the fact that low-dimensional geometric structure exists amongst collections of data.

Linear subspace models are one of the most widely used signal models for collections of high-dimensional data, with applications throughout signal processing, machine learning, and the com-

putational sciences. This is due in part to the simplicity of linear models but also due to the fact that principal components analysis (PCA) provides a closed-form and computationally efficient solution to the problem of finding an optimal low-rank approximation to a collection of data (an ensemble of signals in $\mathbb{R}^n$). More formally, if we stack a collection of d vectors (points) in $\mathbb{R}^n$ into the columns of $Y \in \mathbb{R}^{n \times d}$, then PCA finds the best rank-k estimate of Y by solving

$$\text{(PCA)} \qquad \min_{X \in \mathbb{R}^{n \times d}} \quad \sum_{i=1}^{n} \sum_{j=1}^{d} (Y_{ij} - X_{ij})^2 \quad \text{subject to} \quad \text{rank}(X) \leq k, \tag{1}$$

where X_{ij} is the (i, j) entry of X.

1.1 Unions of Subspaces

In many cases, a linear subspace model is sufficient to characterize the intrinsic structure of an ensemble; however, in many emerging applications, a single subspace is not enough. Instead, ensembles can be modeled as living on a *union of subspaces* or a union of affine planes of mixed or equal dimension. Formally, we say that a set of d signals $\mathcal{Y} = \{y_1, \ldots, y_d\}$, each of dimension n, lives on a union of p subspaces if $\mathcal{Y} \subset \mathcal{U} = \cup_{i=1}^{p} \mathcal{S}_i$, where $\mathcal{S}_i$ is a subspace of $\mathbb{R}^n$.

Ensembles ranging from collections of images taken of objects under different illumination conditions (Basri and Jacobs, 2003; Ramamoorthi, 2002), motion trajectories of point-correspondences (Kanatani, 2001), to structured sparse and block-sparse signals (Lu and Do, 2008; Blumensath and Davies, 2009; Baraniuk et al., 2010) are all well-approximated by a union of low-dimensional subspaces or a union of affine hyperplanes. Union of subspace models have also found utility in the classification of signals collected from complex and adaptive systems at different instances in time, for example, electrical signals collected from the brain's motor cortex (Gowreesunker et al., 2011).

1.2 Exact Feature Selection

Unions of subspaces provide a natural extension to single subspace models, but providing an extension of PCA that leads to provable guarantees for learning multiple subspaces is challenging. This is due to the fact that *subspace clustering*—the identification of points that live in the same subspace—and subspace estimation must be performed simultaneously. However, if we can sift through the points in the ensemble and identify subsets of points that lie along or near the same subspace, then a *local subspace estimate*[1] formed from any such set is guaranteed to coincide with one of the true subspaces present in the ensemble (Vidal et al., 2005; Vidal, 2011). Thus, to guarantee that we obtain an accurate estimate of the subspaces present in a collection of data, we must select a sufficient number of subsets (feature sets) containing points that lie along the same subspace; when a feature set contains points from the same subspace, we say that *exact feature selection* (EFS) occurs.

A common heuristic used for feature selection is to simply select subsets of points that lie within an Euclidean neighborhood of one another (or a fixed number of nearest neighbors (NNs)). Methods that use sets of NNs to learn a union of subspaces include: local subspace affinity (LSA) (Yan and Pollefeys, 2006), spectral clustering based on locally linear approximations (Arias-Castro et al., 2011), spectral curvature clustering (Chen and Lerman, 2009), and local best-fit flats (Zhang et al.,

1. A local subspace estimate is a low-rank approximation formed from a subset of points in the ensemble, rather than from the entire collection of data.

2012). When the subspaces present in the ensemble are non-intersecting and densely sampled, NN-based approaches provide high rates of EFS and in turn, provide accurate local estimates of the subspaces present in the ensemble. However, such approaches quickly fail as the intersection between pairs of subspaces increases and as the number of points in each subspace decreases; in both of these cases, the Euclidean distance between points becomes a poor predictor of whether points belong to the same subspace.

1.3 Endogenous Sparse Recovery

Instead of computing local subspace estimates from sets of NNs, Elhamifar and Vidal (2009) propose a novel approach for feature selection based upon forming sparse representations of the data via ℓ_1-minimization. The main intuition underlying their approach is that when a sparse representation of a point is formed with respect to the remaining points in the ensemble, the representation should only consist of other points that belong to the same subspace. Under certain assumptions on both the sampling and "distance between subspaces",[2] this approach to feature selection leads to provable guarantees that EFS will occur, even when the subspaces intersect (Elhamifar and Vidal, 2010; Soltanolkotabi and Candès, 2012).

We refer to this application of sparse recovery as *endogenous sparse recovery* due to the fact that representations are not formed from an external collection of primitives (such as a basis or dictionary) but are formed "from within" the data. Formally, for a set of d signals $\mathcal{Y} = \{y_1, \ldots, y_d\}$, each of dimension n, the sparsest representation of the i^{th} point $y_i \in \mathbb{R}^n$ is defined as

$$c_i^* \;=\; \arg\min_{c \in \mathbb{R}^d} \|c\|_0 \qquad \text{subject to} \qquad y_i = \sum_{j \neq i} c(j) y_j, \tag{2}$$

where $\|c\|_0$ counts the number of non-zeroes in its argument and $c(j) \in \mathbb{R}$ denotes the contribution of the j^{th} point y_j to the representation of y_i. Let $\Lambda^{(i)} = \text{supp}(c_i^*)$ denote the subset of points selected to represent the i^{th} point and $c_i^*(j)$ denote the contribution of the j^{th} point to the endogenous representation of y_i. By penalizing representations that require a large number of non-zero coefficients, the resulting representation will be sparse.

In general, finding the sparsest representation of a signal has combinatorial complexity; thus, sparse recovery methods such as basis pursuit (BP) (Chen et al., 1998) or low-complexity greedy methods (Davis et al., 1994) are employed to obtain approximate solutions to (2).

1.4 Contributions

In parallel with recent studies of feature selection with ℓ_1-minimization (Elhamifar and Vidal, 2010; Soltanolkotabi and Candès, 2012; Elhamifar and Vidal, 2013), in this paper, we study feature selection with a greedy method for sparse signal recovery known as orthogonal matching pursuit (OMP). The main result of our analysis is a new geometric condition (Theorem 1) for EFS with OMP that highlights the tradeoff between the: *mutual coherence* or similarity between points living in different subspaces and the *covering radius* of the points within the same subspace. The covering radius can be interpreted as the radius of the largest ball that can be embedded within each subspace without touching a point in the ensemble; the vector that lies at the center of this open ball, or the vector in the subspace that attains the covering radius is referred to as a *deep hole*. Theorem 1 suggests that

2. The distance between a pair of subspaces is typically measured with respect to the principal angles between the subspaces or other related distances on the Grassmanian manifold.

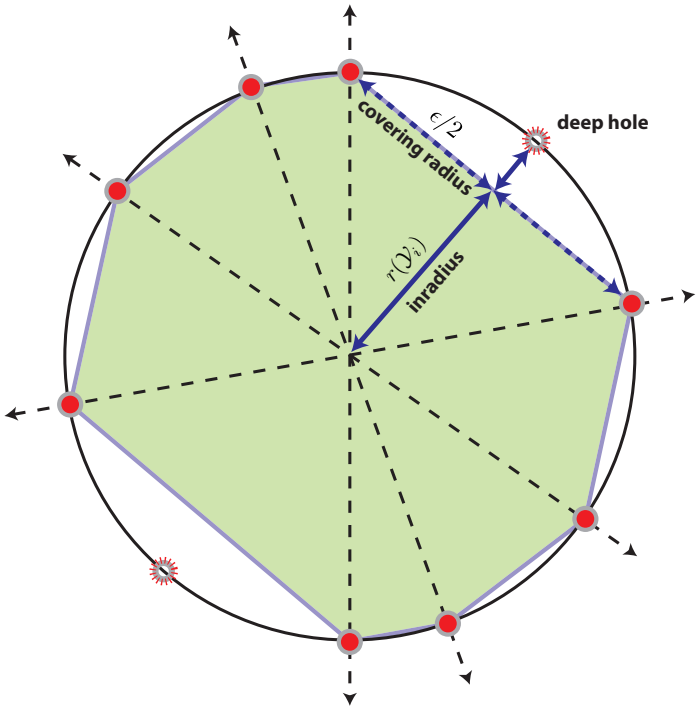

Figure 1: Covering radius of points in a normalized subspace. The interior of the antipodal convex hull of points in a normalized subspace—a subspace of $\mathbb{R}^n$ mapped to the unit ℓ_2-sphere—is shaded. The vector in the normalized subspace (unit circle) that attains the covering radius (deep hole) is marked with a star: when compared with the convex hull, the deep hole coincides with the maximal gap between the convex hull and the set of all vectors that live in the normalized subspace.

subspaces can be arbitrarily close to one another and even intersect, as long as the data is distributed "nicely" along each subspace. By "nicely", we mean that the points that lie on each subspace do not cluster together, leaving large gaps in the sampling of the underlying subspace. In Figure 1, we illustrate the covering radius of a set of points on the sphere (the deep hole is denoted by a star).

After introducing a general geometric condition for EFS, we extend this analysis to the case where the data live on what we refer to as a *bounded union of subspaces* (Theorem 3). In particular, we show that when the points living in a particular subspace are incoherent with the principal vectors that support pairs of subspaces in the ensemble, EFS can be guaranteed, even when non-trivial intersections exist between subspaces in the ensemble. Our condition for bounded subspaces suggests that, in addition to properties related to the sampling of subspaces, one can characterize the separability of pairs of subspaces by examining the correlation between the data set and the unique set of principal vectors that support pairs of subspaces in the ensemble.

In addition to providing a theoretical analysis of EFS with OMP, the other main contribution of this work is revealing the gap between nearest neighbor-based (NN) approaches and sparse recovery methods, that is, OMP and BP, for feature selection. In both synthetic and real world experiments,

we observe that while OMP, BP, and NN provide comparable rates of EFS when subspaces in the ensemble are non-intersecting and densely sampled, sparse recovery methods provide significantly higher rates of EFS than NN sets when: (*i*) the dimension of the intersection between subspaces increases and (*ii*) the sampling density decreases (fewer points per subspace). In Section 5.4, we study the performance of OMP, BP, and NN-based subspace clustering on real data, where the goal is to cluster a collection of images into their respective illumination subspaces. We show that clustering the data with OMP-based feature selection (see Algorithm 2) provides improvements over NN and BP-based (Elhamifar and Vidal, 2010, 2013) clustering methods. In the case of very sparsely sampled subspaces, where the subspace dimension equals 5 and the number of points per subspace equals 16, we obtain a 10% and 30% improvement in classification accuracy with OMP (90%), when compared with BP (80%) and NN (60%).

1.5 Paper Organization

We now provide a roadmap for the rest of the paper.

Section 2. We introduce a signal model for unions of subspaces, detail the sparse subspace clustering (SSC) algorithm (Elhamifar and Vidal, 2010), and then go on to introduce the use of OMP for feature selection and subspace clustering (Algorithm 2); we end with a motivating example.

Section 3 and 4. We provide a formal definition of EFS and then develop the main theoretical results of this paper. We introduce sufficient conditions for EFS to occur with OMP for general unions of subspaces in Theorem 1, disjoint unions in Corollary 1, and bounded unions in Theorem 3.

Section 5. We conduct a number of numerical experiments to validate our theory and compare sparse recovery methods (OMP and BP) with NN-based feature selection. Experiments are provided for both synthetic and real data.

Section 6. We discuss the implications of our theoretical analysis and empirical results on sparse approximation, dictionary learning, and compressive sensing. We conclude with a number of interesting open questions and future lines of research.

Section 7. We supply the proofs of the results contained in Sections 3 and 4.

1.6 Notation

In this paper, we will work solely in real finite-dimensional vector spaces, $\mathbb{R}^n$. We write vectors x in lowercase script, matrices A in uppercase script, and scalar entries of vectors as $x(j)$. The standard p-norm is defined as

$$\|x\|_p = \left(\sum_{j=1}^{n} |x(j)|^p \right)^{1/p},$$

where $p \geq 1$. The "ℓ_0-norm" of a vector x is defined as the number of non-zero elements in x. The support of a vector x, often written as $\text{supp}(x)$, is the set containing the indices of its non-zero coefficients; hence, $\|x\|_0 = |\text{supp}(x)|$. We denote the Moore-Penrose pseudoinverse of a matrix A as $A^\dagger$. If $A = U\Sigma V^T$ then $A^\dagger = V\Sigma^+ U^T$, where we obtain Σ^+ by taking the reciprocal of the entries in Σ, leaving the zeros in their places, and taking the transpose. An orthonormal basis (ONB) Φ that spans the subspace $\mathcal{S}$ of dimension k satisfies the following two properties: $\Phi^T\Phi = I_k$ and $\text{range}(\Phi) = \mathcal{S}$, where I_k is the $k \times k$ identity matrix. Let $P_\Lambda = X_\Lambda X_\Lambda^\dagger$ denote an ortho-projector onto the subspace spanned by the sub-matrix X_Λ.

2. Sparse Feature Selection for Subspace Clustering

In this section, we introduce a signal model for unions of subspaces, detail the sparse subspace clustering (SSC) method (Elhamifar and Vidal, 2009), and introduce an OMP-based method for sparse subspace clustering (SSC-OMP).

2.1 Signal Model

Given a set of p subspaces of $\mathbb{R}^n$, $\{\mathcal{S}_1,\dots,\mathcal{S}_p\}$, we generate a "subspace cluster" by sampling d_i points from the i^{th} subspace $\mathcal{S}_i$ of dimension $k_i \leq k$. Let $\widetilde{\mathcal{Y}}_i$ denote the set of points in the i^{th} subspace cluster and let $\widetilde{\mathcal{Y}} = \cup_{i=1}^{p} \widetilde{\mathcal{Y}}_i$ denote the union of p subspace clusters. Each point in $\widetilde{\mathcal{Y}}$ is mapped to the unit sphere to generate a union of normalized subspace clusters. Let

$$\mathcal{Y} = \left\{ \frac{y_1}{\|y_1\|_2}, \frac{y_2}{\|y_2\|_2}, \cdots, \frac{y_d}{\|y_d\|_2} \right\}$$

denote the resulting set of unit norm points and let $\mathcal{Y}_i$ be the set of unit norm points that lie in the span of subspace $\mathcal{S}_i$. Let $\mathcal{Y}_{-i} = \mathcal{Y} \setminus \mathcal{Y}_i$ denote the set of points in $\mathcal{Y}$ with the points in $\mathcal{Y}_i$ excluded.

Let $Y = [Y_1\ Y_2\ \cdots\ Y_p]$ denote the matrix of normalized data, where each point in $\mathcal{Y}_i$ is stacked into the columns of $Y_i \in \mathbb{R}^{n \times d_i}$. The points in Y_i can be expanded in terms of an ONB $\Phi_i \in \mathbb{R}^{n \times k_i}$ that spans $\mathcal{S}_i$ and subspace coefficients $A_i = \Phi_i^T Y_i$, where $Y_i = \Phi_i A_i$. Let Y_{-i} denote the matrix containing the points in Y with the sub-matrix Y_i excluded.

2.2 Sparse Subspace Clustering

The sparse subspace clustering (SSC) algorithm (Elhamifar and Vidal, 2009) proceeds by solving the following basis pursuit (BP) problem for each point in $\mathcal{Y}$:

$$c_i^* \ = \ \arg\min_{c \in \mathbb{R}^d} \|c\|_1 \qquad \text{subject to} \qquad y_i = \sum_{j \neq i} c(j) y_j.$$

After solving BP for each point in the ensemble, each d-dimensional feature vector c_i^* is placed into the i^{th} row or column of a matrix $C \in \mathbb{R}^{d \times d}$ and spectral clustering (Shi and Malik, 2000; Ng et al., 2002) is performed on the graph Laplacian of the affinity matrix $W = |C| + |C^T|$.

In situations where points might not admit an exact representation with respect to other points in the ensemble, an inequality constrained version of BP known as basis pursuit denoising (BPDN) may be employed for feature selection (Elhamifar and Vidal, 2013). In this case, the following BPDN problem is computed for each point in $\mathcal{Y}$:

$$c_i^* \ = \ \arg\min_{c \in \mathbb{R}^d} \|c\|_1 \qquad \text{subject to} \qquad \|y_i - \sum_{j \neq i} c(j) y_j\|_2 < \kappa, \tag{3}$$

where κ is a parameter that is selected based upon the amount of noise in the data. Recently, Wang and Xu (2013) provided an analysis of EFS for a unconstrained variant of the formulation in (3) for noisy unions of subspaces. Soltanolkotabi et al. (2013) proposed a robust procedure for subspace clustering from noisy data that extends the BPDN framework studied in Elhamifar and Vidal (2013) and Wang and Xu (2013).

Algorithm 1 : **Orthogonal Matching Pursuit (OMP)**

Input: Input signal $y \in \mathbb{R}^n$, a matrix $A \in \mathbb{R}^{n \times d}$ containing d signals $\{a_i\}_{i=1}^d$ in its columns, and a stopping criterion (either the sparsity k or the approximation error κ).

Output: An index set Λ containing the indices of all atoms selected in the pursuit and a coefficient vector c containing the coefficients associated with of all atoms selected in the pursuit.

Initialize: Set the residual to the input signal $s = y$.

1. Select the atom that is maximally correlated with the residual and add it to Λ

$$\Lambda \leftarrow \Lambda \cup \arg\max_i |\langle a_i, s \rangle|.$$

2. Update the residual by projecting s into the space orthogonal to the span of A_Λ

$$s \leftarrow (I - P_\Lambda) y.$$

3. Repeat steps (1)–(2) until the stopping criterion is reached, for example, either $|\Lambda| = k$ or the norm of the residual $\|s\|_2 \leq \kappa$.
4. Return the support set Λ and coefficient vector $c = A_\Lambda^\dagger y$.

2.3 Greedy Feature Selection

Instead of solving the sparse recovery problem in (2) via ℓ_1-minimization, we propose the use of a low-complexity method for sparse recovery known as orthogonal matching pursuit (OMP). We detail the OMP algorithm in Algorithm 1. For each point y_i, we solve Algorithm 1 to obtain a sparse representation of the signal with respect to the remaining points in Y. The output of the OMP algorithm is a *feature set*, $\Lambda^{(i)}$, which indexes the columns in Y selected to form an endogenous representation of y_i.

After computing feature sets for each point in the data set via OMP, either a spectral clustering method or a consensus-based method (Dyer, 2011) may then be employed to cluster the data. In Algorithm 2, we outline a procedure for performing an OMP-based variant of the SSC algorithm that we will refer to as SSC-OMP.

2.4 Motivating Example: Clustering Illumination Subspaces

As a motivating example, we consider the problem of clustering a collection of images of faces captured under different illumination conditions: images of Lambertian objects (no specular reflections) captured under different illumination conditions have been shown to be well-approximated by a 5-dimensional subspace (Ramamoorthi, 2002). In Figure 2, we show an example of the affinity matrices obtained via OMP, BP, and NN, for a collection of images of two faces under 64 different illumination conditions selected from the Yale B Database (Georghiades et al., 2001). In this example, each $N \times N$ image is considered a point in $\mathbb{R}^{N^2}$ and the images of a particular person's face captured under different illumination conditions lie on a low-dimensional subspace. By varying the number of unique illumination conditions that we collect (number of points per subspace), we can manipulate the sampling density of the subspaces in the ensemble. We sort the images (points) such that images of the same face are contained in a contiguous block.

Algorithm 2 : **Sparse Subspace Clustering with OMP (SSC-OMP)**

Input: A data set $Y \in \mathbb{R}^{n \times d}$ containing d points $\{y_i\}_{i=1}^{d}$ in its columns, a stopping criterion for OMP, and the number of clusters p.

Output: A set of d labels $\mathcal{L} = \{\ell(1), \ell(2), \ldots, \ell(d)\}$, where $\ell(i) \in \{1, 2, \ldots, p\}$ is the label associated with the i^{th} point y_i.

Step 1. Compute Subspace Affinity via OMP

1. Solve Algorithm 1 for the i^{th} point y_i to obtain a feature set Λ and coefficient vector c.
2. For all $j \in \Lambda^{(i)}$, let $C_{ij} = c(j)$. Otherwise, set $C_{ij} = 0$.
3. Repeat steps (1)–(2) for all $i = 1, \ldots, d$.

Step 2. Perform Spectral Clustering

1. Symmetrize the subspace affinity matrix C to obtain $W = |C| + |C^T|$.
2. Perform spectral clustering on W to obtain a set of d labels $\mathcal{L}$.

To generate the OMP affinity matrices in the right column, we use the greedy feature selection procedure outlined in Step 1 of Algorithm 2, where the sparsity level $k = 5$. To generate the BP affinity matrices in the middle column, we solved the BPDN problem in (3) via a homotopy algorithm where we vary the noise parameter κ and choose the smallest value of κ that produces up to 5 coefficients. The resulting coefficient vectors are then stacked into the rows of a matrix C and the final subspace affinity W is computed by symmetrizing the coefficient matrix, $W = |C| + |C^T|$. To generate the NN affinity matrices in the left column, we compute the absolute normalized inner products between all points in the data set and then threshold each row to select the $k = 5$ nearest neighbors to each point.

3. Geometric Analysis of Exact Feature Selection

In this section, we provide a formal definition of EFS and develop sufficient conditions that guarantee that EFS will occur for all of the points contained within a particular subspace cluster.

3.1 Exact Feature Selection

In order to guarantee that OMP returns a feature set (subset of points from $\mathcal{Y}$) that produces an accurate local subspace estimate, we will be interested in determining when the feature set returned by Algorithm 1 only contains points that belong to the same subspace cluster, that is, *exact feature selection* (EFS) occurs. EFS provides a natural condition for studying performance of both subspace consensus and spectral clustering methods due to the fact that when EFS occurs, this results in a local subspace estimate that coincides with one of the true subspaces contained within the data. We now supply a formal definition of EFS.

Definition 1 (Exact Feature Selection) *Let $\mathcal{Y}_k = \{y : (I - P_k)y = 0,\ y \in \mathcal{Y}\}$ index the set of points in $\mathcal{Y}$ that live in the span of subspace $\mathcal{S}_k$, where P_k is a projector onto the span of subspace $\mathcal{S}_k$. For a point $y \in \mathcal{Y}_k$ with feature set Λ, if $y_i \subseteq \mathcal{Y}_k$, $\forall i \in \Lambda$, we say that Λ contains exact features.*

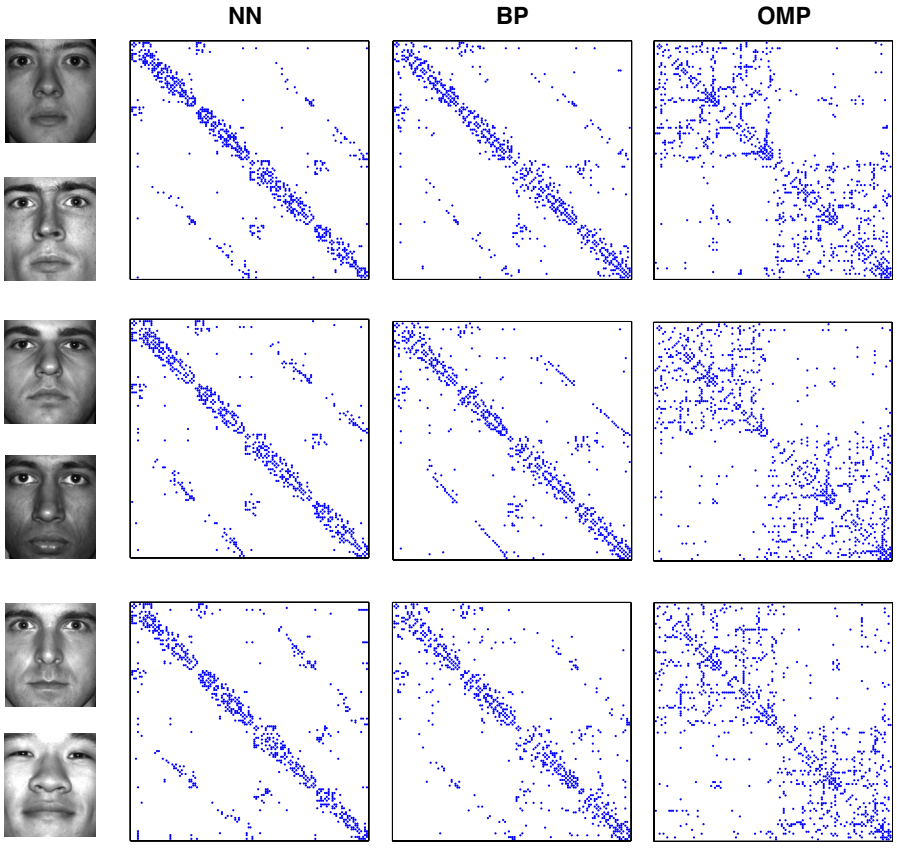

Figure 2: Comparison of subspace affinity matrices for illumination subspaces. In each row, we display the affinity matrices obtained for a different pair of illumination subspaces, for NN (left), BP (middle), and OMP (right). To the left of the affinity matrices, we display an exemplar image from each illumination subspace.

3.2 Sufficient Conditions for EFS

In this section, we develop geometric conditions that are sufficient for EFS with OMP. Before proceeding, however, we must introduce properties of the data set required to develop our main results.

3.2.1 PRELIMINARIES

Our main geometric result in Theorem 1 below requires measures of both the distance between points in *different subspace clusters* and within the *same subspace cluster*. A natural measure of the similarity between points living in different subspaces is the *mutual coherence*. A formal definition of the mutual coherence is provided below in Def. 2.

Definition 2 (Mutual Coherence) *The mutual coherence between the points in the sets* $(\mathcal{Y}_i, \mathcal{Y}_j)$ *is defined as*

$$\mu_c(\mathcal{Y}_i, \mathcal{Y}_j) \;=\; \max_{u \in \mathcal{Y}_i, v \in \mathcal{Y}_j} |\langle u, v \rangle|, \text{ where } \|u\|_2 = \|v\|_2 = 1.$$

In words, the mutual coherence provides a point-wise measure of the normalized inner product (coherence) between all pairs of points that lie in two different subspace clusters.

Let $\mu_c(\mathcal{Y}_i)$ denote the maximum mutual coherence between the points in $\mathcal{Y}_i$ and all other subspace clusters in the ensemble, where

$$\mu_c(\mathcal{Y}_i) = \max_{j \neq i} \mu_c(\mathcal{Y}_i, \mathcal{Y}_j).$$

A related quantity that provides an upper bound on the mutual coherence is the cosine of the first *principal angle* between the subspaces. The first principal angle θ^*_{ij} between subspaces $\mathcal{S}_i$ and $\mathcal{S}_j$, is the smallest angle between a pair of unit vectors (u_1, v_1) drawn from $\mathcal{S}_i \times \mathcal{S}_j$. Formally, the first principal angle is defined as

$$\theta^*_{ij} = \min_{u \in \mathcal{S}_i,\ v \in \mathcal{S}_j} \arccos \langle u, v \rangle \quad \text{subject to} \quad \|u\|_2 = 1, \|v\|_2 = 1. \tag{4}$$

Whereas the mutual coherence provides a measure of the similarity between a pair of unit norm vectors that are contained in the sets $\mathcal{Y}_i$ and $\mathcal{Y}_j$, the cosine of the minimum principal angle provides a measure of the similarity between all pairs of unit norm vectors that lie in the span of $\mathcal{S}_i \times \mathcal{S}_j$. For this reason, the cosine of the first principal angle provides an upper bound on the mutual coherence. The following upper bound is in effect for each pair of subspace clusters in the ensemble:

$$\mu_c(\mathcal{Y}_i, \mathcal{Y}_j) \leq \cos(\theta^*_{ij}). \tag{5}$$

To measure how well points in the same subspace cluster cover the subspace they live on, we will study the covering radius of each normalized subspace cluster relative to the projective distance. Formally, the covering radius of the set $\mathcal{Y}_k$ is defined as

$$\text{cover}(\mathcal{Y}_k) \ = \ \max_{u \in \mathcal{S}_k} \min_{y \in \mathcal{Y}_k} \text{dist}(u, y),$$

where the projective distance between two vectors u and y is defined relative to the acute angle between the vectors

$$\text{dist}(u, y) = \sqrt{1 - \frac{|\langle u, y \rangle|^2}{\|u\|_2 \|y\|_2}}.$$

The covering radius of the normalized subspace cluster $\mathcal{Y}_i$ can be interpreted as the size of the largest open ball that can be placed in the set of all unit norm vectors that lie in the span of $\mathcal{S}_i$, without touching a point in $\mathcal{Y}_i$.

Let (u^*_i, y^*_i) denote a pair of points that attain the maximum covering diameter for $\mathcal{Y}_i$; u^*_i is referred to as a deep hole in $\mathcal{Y}_i$ along $\mathcal{S}_i$. The covering radius can be interpreted as the sine of the angle between the deep hole $u^*_i \in \mathcal{S}_i$ and its nearest neighbor $y^*_i \in \mathcal{Y}_i$. We show the geometry underlying the covering radius in Figure 1.

In the sequel, we will be interested in the maximum (worst-case) covering attained over all d_i sets formed by removing a single point from $\mathcal{Y}_i$. We supply a formal definition below in Def. 3.

Definition 3 (Covering Radius) *The maximum covering diameter* ε *of the set* $\mathcal{Y}_i$ *along the subspace* $\mathcal{S}_i$ *is defined as*

$$\varepsilon \ = \ \max_{j=1,\ldots,d_i} 2 \ \text{cover}(\{\mathcal{Y}_i \setminus y_j\}).$$

Hence, the covering radius equals $\varepsilon/2$.

A related quantity is the *inradius* of the set $\mathcal{Y}_i$, or the cosine of the angle between a point in $\mathcal{Y}_i$ and any point in $\mathcal{S}_i$ that attains the covering radius. The relationship between the covering diameter ε and inradius $r(\mathcal{Y}_i)$ is given by

$$r(\mathcal{Y}_i) = \sqrt{1 - \frac{\varepsilon^2}{4}}. \tag{6}$$

A geometric interpretation of the inradius is that it measures the distance from the origin to the maximal gap in the antipodal convex hull of the points in $\mathcal{Y}_i$. The geometry underlying the covering radius and the inradius is displayed in Figure 1.

3.2.2 MAIN RESULT FOR EFS

We are now equipped to state our main result for EFS with OMP. The proof is contained in Section 7.1.

Theorem 1 *Let ε denote the maximal covering diameter of the subspace cluster $\mathcal{Y}_i$ as defined in Def. 3. A sufficient condition for Algorithm 1 to return exact feature sets for all points in $\mathcal{Y}_i$ is that the mutual coherence*

$$\mu_c(\mathcal{Y}_i) \;<\; \sqrt{1 - \frac{\varepsilon^2}{4}} - \frac{\varepsilon}{\sqrt[4]{12}} \max_{j \neq i} \cos(\theta^*_{ij}), \tag{7}$$

where θ^*_{ij} is the minimum principal angle defined in (4).

In words, this condition requires that the mutual coherence between points in *different subspaces* is less than the difference of two terms that both depend on the covering radius of points along a *single subspace*. The first term on the RHS of (7) is equal to the inradius, as defined in (6). The second term on the RHS of (7) is the product of the cosine of the minimum principal angle between pairs of subspaces in the ensemble and the covering diameter ε of the points in $\mathcal{Y}_i$.

When subspaces in the ensemble intersect, that is, $\cos(\theta^*_{ij}) = 1$, condition (7) in Theorem 1 can be simplified as

$$\mu_c(\mathcal{Y}_i) < \sqrt{1 - \frac{\varepsilon^2}{4}} - \frac{\varepsilon}{\sqrt[4]{12}} \approx \sqrt{1 - \frac{\varepsilon^2}{4}} - \frac{\varepsilon}{1.86}.$$

In this case, EFS can be guaranteed for intersecting subspaces as long as the points in distinct subspace clusters are bounded away from intersections between subspaces. When the covering radius shrinks to zero, Theorem 1 requires that $\mu_c < 1$, or that points from different subspaces do not lie exactly in the subspace intersection, that is, are identifiable from one another.

3.2.3 EFS FOR DISJOINT SUBSPACES

When the subspaces in the ensemble are *disjoint*, that is, $\cos(\theta^*_{ij}) < 1$, Theorem 1 can be simplified further by using the bound for the mutual coherence in (5). This simplification results in the following corollary.

Corollary 1 *Let θ^*_{ij} denote the first principal angle between a pair of disjoint subspaces $\mathcal{S}_i$ and $\mathcal{S}_j$, and let ε denote the maximal covering diameter of the points in $\mathcal{Y}_i$. A sufficient condition for Algorithm 1 to return exact feature sets for all points in $\mathcal{Y}_i$ is that*

$$\max_{j \neq i} \cos(\theta^*_{ij}) \;<\; \frac{\sqrt{1 - \varepsilon^2/4}}{1 + \varepsilon/\sqrt[4]{12}}.$$

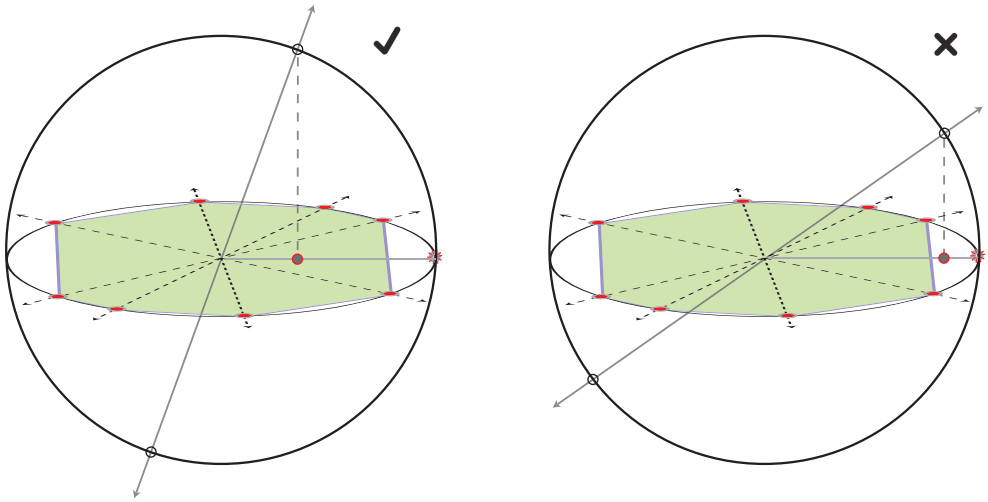

Figure 3: Geometry underlying EFS. A union of two disjoint subspaces of different dimension: the convex hull of a set of points (red circles) living on a 2D subspace is shaded (green). In (a), we show an example where EFS is guaranteed—the projection of points along the 1D subspace lie inside the shaded region. In (b), we show an example where EFS is not guaranteed—the projection of points along the 1D subspace lie outside the shaded region.

3.2.4 GEOMETRY UNDERLYING EFS WITH OMP

The main idea underlying the proof of Theorem 1 is that, at each iteration of Algorithm 1, we require that the residual used to select a point to be included in the feature set is closer to a point in the *correct subspace cluster* ($\mathcal{Y}_i$) than a point in an *incorrect subspace cluster* ($\mathcal{Y}_{-i}$). To be precise, we require that the normalized inner product of the residual signal s and all points outside of the correct subspace cluster

$$\max_{y \in \mathcal{Y}_{-i}} \frac{|\langle s, y \rangle|}{\|s\|_2} < r(\mathcal{Y}_i), \tag{8}$$

at each iteration of Algorithm 1. To provide the result in Theorem 1, we require that (8) holds for all $s \in \mathcal{S}_i$, or all possible residual vectors.

A geometric interpretation of the EFS condition in Theorem 1 is that the orthogonal projection of all points outside of a subspace must lie within the antipodal convex hull of the set of normalized points that span the subspace. To see this, consider the projection of the points in $\mathcal{Y}_{-i}$ onto $\mathcal{S}_i$. Let z_j^* denote the point on subspace $\mathcal{S}_i$ that is closest to the signal $y_j \in \mathcal{Y}_{-i}$,

$$z_j^* = \arg\min_{z \in \mathcal{S}_i} \|z - y_j\|_2.$$

We can also write this projection in terms of a orthogonal projection operator $P_i = \Phi_i \Phi_i^T$, where Φ_i is an ONB that spans $\mathcal{S}_i$ and $z_j^* = P_i y_j$.

By definition, the normalized inner product of the residual with points in incorrect subspace clusters is upper bounded as

$$\max_{y_j \in \mathcal{Y}_{-i}} \frac{|\langle s, y_j \rangle|}{\|s\|_2} \leq \max_{y_j \in \mathcal{Y}_{-i}} \frac{|\langle z_j^*, y_j \rangle|}{\|z_j^*\|_2} = \max_{y_j \in \mathcal{Y}_{-i}} \cos \angle \{z_j^*, y_j\}$$

Thus to guarantee EFS, we require that the cosine of the angle between all signals in $\mathcal{Y}_{-i}$ and their projection onto $\mathcal{S}_i$ is less than the inradius of $\mathcal{Y}_i$. Said another way, the EFS condition requires that the length of all projected points be less than the inradius of $\mathcal{Y}_i$.

In Figure 3, we provide a geometric visualization of the EFS condition for a union of disjoint subspaces (union of a 1D subspace with a 2D subspace). In (a), we show an example where EFS is guaranteed because the projection of the points outside of the 2D subspace lie well within the antipodal convex hull of the points along the normalized 2D subspace (ring). In (b), we show an example where EFS is not guaranteed because the projection of the points outside of the 2D subspace lie outside of the antipodal convex hull of the points along the normalized 2D subspace (ring).

3.3 Connections to Previous Work

In this section, we will connect our results for OMP with previous analyses of EFS with BP for disjoint (Elhamifar and Vidal, 2010, 2013) and intersecting (Soltanolkotabi and Candès, 2012) subspaces. Following this, we will contrast the geometry underlying EFS with exact recovery conditions used to guarantee support recovery for both OMP and BP (Tropp, 2004, 2006).

3.3.1 SUBSPACE CLUSTERING WITH BP

Elhamifar and Vidal (2010) develop the following sufficient condition for EFS to occur for BP from a union of disjoint subspaces,

$$\max_{j \neq i} \cos(\theta_{ij}^*) \; < \; \max_{\widetilde{Y}_i \in \mathbb{W}_i} \frac{\sigma_{\min}(\widetilde{Y}_i)}{\sqrt{k_i}}, \tag{9}$$

where $\mathbb{W}_i$ is the set of all full rank sub-matrices $\widetilde{Y}_i \in \mathbb{R}^{n \times k_i}$ of the data matrix $Y_i \in \mathbb{R}^{n \times d_i}$ and $\sigma_{\min}(\widetilde{Y}_i)$ is the minimum singular value of the sub-matrix $\widetilde{Y}_i$. Since we assume that all of the data points have been normalized, $\sigma_{\min}(\widetilde{Y}_i) \leq 1$; thus, the best case result that can be obtained is that the minimum principal angle, $\cos(\theta_{ij}^*) < 1/\sqrt{k_i}$. This suggests that the minimum principal angle of the union must go to zero, that is, the union must consist of orthogonal subspaces, as the subspace dimension increases.

In contrast to the condition in (9), the conditions we provide in Theorem 1 and Corollary 1 do not depend on the subspace dimension. Rather, we require that there are enough points in each subspace to achieve a sufficiently small covering; in which case, EFS can be guaranteed for subspaces of any dimension.

Soltanolkotabi and Candès (2012) develop the following sufficient condition for EFS to occur for BP from a union of intersecting subspaces,

$$\mu_v(\mathcal{Y}_i) = \max_{y \in \mathcal{Y}_{-i}} \|V_{(i)}{}^T y\|_\infty < r(\mathcal{Y}_i), \tag{10}$$

where the matrix $V_{(i)} \in \mathbb{R}^{d_i \times n}$ contains the dual directions (the dual vectors for each point in $\mathcal{Y}_i$ embedded in $\mathbb{R}^n$) in its columns,[3] and $r(\mathcal{Y}_i)$ is the inradius as defined in (6). In words, (10) requires that the maximum coherence between any point in $\mathcal{Y}_{-i}$ and the dual directions contained in $V_{(i)}$ be less than the inradius of the points in $\mathcal{Y}_i$.

To link the result in (10) to our guarantee for OMP in Theorem 1, we observe that while (10) requires that $\mu_v(\mathcal{Y}_i)$ (coherence between a point in a subspace cluster and the dual directions of points in a different subspace cluster) be less than the inradius, Theorem 1 requires that the mutual coherence $\mu_c(\mathcal{Y}_i)$ (coherence between two points in different subspace clusters) be less than the inradius minus an additional term that depends on the covering radius. For an arbitrary set of points that live on a union of subspaces, the precise relationship between the two coherence parameters $\mu_c(\mathcal{Y}_i)$ and $\mu_v(\mathcal{Y}_i)$ is not straightforward; however, when the points in each subspace cluster are distributed uniformly and at random along each subspace, the dual directions will also be distributed uniformly along each subspace.[4] In this case, $\mu_v(\mathcal{Y}_i)$ will be roughly equivalent to the mutual coherence $\mu_c(\mathcal{Y}_i)$.

This simplification reveals the connection between the result in (10) for BP and the condition in Theorem 1 for OMP. In particular, when $\mu_v(\mathcal{Y}_i) \approx \mu_c(\mathcal{Y}_i)$, our result for OMP requires that the mutual coherence is smaller than the inradius minus an additional term that is linear in the covering diameter ε. For this reason, our result in Theorem 1 is more restrictive than the result provided in (10). The gap between the two bounds shrinks to zero only when the minimum principal angle $\theta^*_{ij} \to \pi/2$ (orthogonal subspaces) or when the covering diameter $\varepsilon \to 0$.

In our empirical studies, we find that when BPDN is tuned to an appropriate value of the noise parameter κ, BPDN tends to produce higher rates of EFS than OMP. This suggests that the theoretical gap between the two approaches might not be an artifact of our current analysis; rather, there might exist an intrinsic gap between the performance of each method with respect to EFS. Nonetheless, an interesting finding from our empirical study in Section 5.4, is that despite the fact that BPDN provides better rates of EFS than OMP, OMP typically provides better clustering results than BPDN. For these reasons, we maintain that OMP offers a powerful low-complexity alternative to ℓ_1-minimization approaches for feature selection.

3.3.2 EXACT RECOVERY CONDITIONS FOR SPARSE RECOVERY

To provide further intuition about EFS in endogenous sparse recovery, we will compare the geometry underlying the EFS condition with the geometry of the exact recovery condition (ERC) for sparse signal recovery methods (Tropp, 2004, 2006).

To guarantee exact support recovery for a signal $y \in \mathbb{R}^n$ which has been synthesized from a linear combination of atoms from the sub-matrix $\Phi_\Lambda \in \mathbb{R}^{n \times k}$, we must ensure that our approximation of y consists solely of atoms from Φ_Λ. Let $\{\varphi_i\}_{i \notin \Lambda}$ denote the set of atoms in Φ that are not indexed by the set Λ. The *exact recovery condition* (ERC) in Theorem 2 is sufficient to guarantee that we obtain exact support recovery for both BP and OMP (Tropp, 2004).

3. See Def. 2.2 for a formal definition of the dual directions and insight into the geometry underlying their guarantees for EFS via BP (Soltanolkotabi and Candès, 2012).
4. This approximation is based upon personal correspondence with M. Soltankotabi, an author of the work in Soltanolkotabi and Candès (2012).

Theorem 2 (Tropp, 2004) *For any signal supported over the sub-dictionary Φ_Λ, exact support recovery is guaranteed for both OMP and BP if*

$$\mathrm{ERC}(\Lambda) = \max_{i \notin \Lambda} \|\Phi_\Lambda{}^\dagger \varphi_i\|_1 < 1.$$

A geometric interpretation of the ERC is that it provides a measure of how far a projected atom φ_i outside of the set Λ lies from the antipodal convex hull of the atoms in Λ. When a projected atom lies outside of the antipodal convex hull formed by the set of points in the sub-dictionary Φ_Λ, then the ERC condition is violated and support recovery is not guaranteed. For this reason, the ERC requires that the maximum coherence between the atoms in Φ is sufficiently low or that Φ is *incoherent*.

While the ERC condition requires a *global incoherence* property on all of the columns of Φ, we can interpret EFS as requiring a *local incoherence* property. In particular, the EFS condition requires that the projection of atoms in an incorrect subspace cluster $\mathcal{Y}_{-i}$ onto $\mathcal{S}_i$ must be incoherent with any deep holes in $\mathcal{Y}_i$ along $\mathcal{S}_i$. In contrast, we require that the points within a subspace cluster exhibit local coherence in order to produce a small covering radius.

4. EFS for Bounded Unions of Subspaces

In this section, we study the connection between EFS and the higher-order principal angles (beyond the minimum angle) between pairs of intersecting subspaces.

4.1 Subspace Distances

To characterize the "distance" between pairs of subspaces in the ensemble, the *principal angles* between subspaces will prove useful. As we saw in the previous section, the first principal angle θ_0 between subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ of dimension k_1 and k_2 is defined as the smallest angle between a pair of unit vectors (u_1, v_1) drawn from $\mathcal{S}_1 \times \mathcal{S}_2$. The vector pair (u_1^*, v_1^*) that attains this minimum is referred to as the first set of principal vectors. The second principal angle θ_1 is defined much like the first, except that the second set of principal vectors that define the second principal angle are required to be orthogonal to the first set of principal vectors (u_1^*, v_1^*). The remaining principal angles are defined recursively in this way. The sequence of $k = \min(k_1, k_2)$ principal angles, $\theta_0 \leq \theta_1 \leq \cdots \leq \theta_{k-1}$, is non-decreasing and all of the principal angles lie between $[0, \pi/2]$.

The definition above provides insight into what the principal angles/vectors tell us about the geometry underlying a pair of subspaces; in practice, however, the principal angles are not computed in this recursive manner. Rather, a computationally efficient way to compute the principal angles between two subspaces $\mathcal{S}_i$ and $\mathcal{S}_j$ is to first compute the singular values of the matrix $G = \Phi_i^T \Phi_j$, where $\Phi_i \in \mathbb{R}^{n \times k_i}$ is an ONB that spans subspace $\mathcal{S}_i$. Let $G = U \Sigma V^T$ denote the SVD of G and let $\sigma_{ij} \in [0,1]^k$ denote the singular values of G, where $k = \min(k_i, k_j)$ is the minimum dimension of the two subspaces. The m^{th} smallest principal angle $\theta_{ij}(m)$ is related to the m^{th} largest entry of σ_{ij} via the following relationship, $\cos(\theta_{ij}(m)) = \sigma_{ij}(m)$. For our subsequent discussion, we will refer to the singular values of G as the *cross-spectra* of the subspace pair $(\mathcal{S}_i, \mathcal{S}_j)$.

A pair of subspaces is said to be *disjoint* if the minimum principal angle is greater than zero. Non-disjoint or intersecting subspaces are defined as subspaces with minimum principal angle equal to zero. The dimension of the intersection between two subspaces is equivalent to the number of principal angles equal to zero or equivalently, the number of entries of the cross-spectra that

are equal to one. We define the *overlap* between two subspaces as the $\text{rank}(G)$ or equivalently, $q = \|\sigma_{ij}\|_0$, where $q \geq \dim(\mathcal{S}_i \cap \mathcal{S}_j)$.

4.2 Sufficient Conditions for EFS from Bounded Unions

The sufficient conditions for EFS in Theorem 1 and Corollary 1 reveal an interesting relationship between the covering radius, mutual coherence, and the minimum principal angle between pairs of subspaces in the ensemble. However, we have yet to reveal any dependence between EFS and higher-order principal angles. To make this connection more apparent, we will make additional assumptions about the distribution of points in the ensemble, namely that the data set produces a *bounded union of subspaces* relative to the principal vectors supporting pairs of subspaces in the ensemble.

Let $Y = [Y_i\, Y_j]$ denote a collection of unit-norm data points, where Y_i and Y_j contain the points in subspaces $\mathcal{S}_i$ and $\mathcal{S}_j$, respectively. Let $G = \Phi_i^T \Phi_j = U \Sigma V^T$ denote the SVD of G, where $\text{rank}(G) = q$. Let $\widetilde{U} = \Phi_i U_q$ denote the set of left principal vectors of G that are associated with the q nonzero singular values in Σ. Similarly, let $\widetilde{V} = \Phi_j V_q$ denote the set of right principal vectors of G that are associated with the nonzero singular values in Σ. When the points in each subspace are incoherent with the principal vectors in the columns of $\widetilde{U}$ and $\widetilde{V}$, we say that the ensemble Y is an *bounded union of subspaces*. Formally, we require the following incoherence property holds:

$$\left(\|Y_i^T \widetilde{U}\|_\infty, \|Y_j^T \widetilde{V}\|_\infty \right) \leq \gamma, \tag{11}$$

where $\|\cdot\|_\infty$ is the entry-wise maximum and $\gamma \in (0,1]$. This property requires that the inner products between the points in a subspace and the set of principal vectors that span non-orthogonal directions between a pair of subspaces is bounded by a fixed constant.

When the points in each subspace are distributed such that (11) holds, we can rewrite the mutual coherence between any two points from different subspaces to reveal its dependence on higher-order principal angles. In particular, we show (in Section 7.2) that the coherence between the residual s used in Algorithm 1 to select the next point to be included in the representation of a point $y \in \mathcal{Y}_i$, and a point in $\mathcal{Y}_j$ is upper bounded by

$$\max_{y \in \mathcal{Y}_j} \frac{|\langle s, y \rangle|}{\|s\|_2} \leq \gamma \|\sigma_{ij}\|_1, \tag{12}$$

where γ is the bounding constant of the data Y and $\|\sigma_{ij}\|_1$ is the ℓ_1-norm of the cross-spectra or equivalently, the trace norm of G. Using the bound in (12), we arrive at the following sufficient condition for EFS from bounded unions of subspaces. We provide the proof in Section 7.2.

Theorem 3 *Let Y live on a bounded union of subspaces, where $q = \text{rank}(G)$ and $\gamma < \sqrt{1/q}$. Let σ_{ij} denote the cross-spectra of the subspaces $\mathcal{S}_i$ and $\mathcal{S}_j$ and let ε denote the covering diameter of $\mathcal{Y}_i$. A sufficient condition for Algorithm 1 to return exact feature sets for all points in $\mathcal{Y}_i$ is that the covering diameter*

$$\varepsilon \;<\; \min_{j \neq i} \sqrt{1 - \gamma^2 \|\sigma_{ij}\|_1^2}.$$

This condition requires that both the covering diameter of each subspace and the bounding constant of the union be sufficiently small in order to guarantee EFS. One way to guarantee that the

ensemble has a small bounding constant is to constrain the total amount of energy that points in $\mathcal{Y}_j$ have in the q-dimensional subspace spanned by the principal vectors in $\widetilde{V}$.

Our analysis for bounded unions assumes that the nonzero entries of the cross-spectra are equal, and thus each pair of supporting principal vectors in $\widetilde{V}$ are equally important in determining whether points in $\mathcal{Y}_i$ will admit EFS. However, this assumption is not true in general. When the union is supported by principal vectors with non-uniform principal angles, our analysis suggests that a weaker form of incoherence is required. Instead of requiring incoherence with all principal vectors, the data must be sufficiently incoherent with the principal vectors that correspond to small principal angles (or large values of the cross-spectra). This means that as long as points are not concentrated along the principal directions with small principal angles (i.e., intersections), then EFS can be guaranteed, even when subspaces exhibit non-trivial intersections. To test this prediction, we will study EFS for a *bounded energy model* in Section 5.2. We show that when the data set is sparsely sampled (larger covering radius), reducing the amount of energy that points contain in subspace intersections, does in fact increase the probability that points admit EFS.

Finally, our analysis of bounded unions suggests that the decay of the cross-spectra is likely to play an important role in determining whether points will admit EFS or not. To test this hypothesis, we will study the role that the structure of the cross-spectra plays in EFS in Section 5.3.

5. Experimental Results

In our theoretical analysis of EFS in Sections 3 and 4, we revealed an intimate connection between the covering radius of subspaces and the principal angles between pairs of subspaces in the ensemble. In this section, we will conduct an empirical study to explore these connections further. In particular, we will study the probability of EFS as we vary the covering radius as well as the dimension of the intersection and/or overlap between subspaces.

5.1 Generative Model for Synthetic Data

In order to study EFS for unions of subspaces with varied cross-spectra, we will generate synthetic data from unions of overlapping *block sparse signals*.

5.1.1 CONSTRUCTING SUB-DICTIONARIES

We construct a pair of sub-dictionaries as follows: Take two subsets Ω_1 and Ω_2 of k signals (atoms) from a dictionary D containing M atoms $\{d_m\}_{m=1}^M$ in its columns, where $d_m \in \mathbb{R}^n$ and $|\Omega_1| = |\Omega_2| = k$. Let $\Psi \in \mathbb{R}^{n\times k}$ denote the subset of atoms indexed by Ω_1, and let $\Phi \in \mathbb{R}^{n\times k}$ denote the subset of atoms indexed by Ω_2. Our goal is to select Ψ and Φ such that $G = \Psi^T\Phi$ is diagonal, that is, $\langle \psi_i, \phi_j \rangle = 0$, if $i \neq j$, where ψ_i is the i^{th} element in Ψ and ϕ_j is the j^{th} element of Φ. In this case, the cross-spectra is defined as $\sigma = \text{diag}(G)$, where $\sigma \in [0,1]^k$. For each union, we fix the "overlap" q or the rank of $G = \Psi^T\Phi$ to a constant between zero (orthogonal subspaces) and k (maximal overlap).

To generate a pair of k-dimensional subspaces with a q-dimensional overlap, we can pair the elements from Ψ and Φ such that the i^{th} entry of the cross-spectra equals

$$\sigma(i) = \begin{cases} |\langle \psi_i, \phi_i \rangle| & \text{if } 1 \leq i \leq q, \\ 0 & \text{if } i = q+1 \leq i \leq k. \end{cases}$$

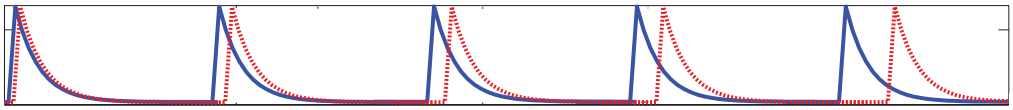

Figure 4: Generating unions of subspaces from shift-invariant dictionaries. An example of a collection of two sub-dictionaries of five atoms, where each of the atoms have a non-zero inner product with one other atom. This choice of sub-dictionaries produces a union of disjoint subspaces, where the overlap ratio $\delta = q/k = 1$.

We can leverage the banded structure of shift-invariant dictionaries, for example, dictionary matrices with localized Toeplitz structure, to generate subspaces with structured cross-spectra as follows.[5] First, we fix a set of k incoherent (orthogonal) atoms from our shift-invariant dictionary, which we place in the columns of Ψ. Now, holding Ψ fixed, we set the i^{th} atom ϕ_i of the second sub-dictionary Φ to be a shifted version of the i^{th} atom ψ_i of the dictionary Ψ. To be precise, if we set $\psi_i = d_m$, where d_m is the m^{th} atom in our shift-invariant dictionary, then we will set $\phi_i = d_{m+\Delta}$ for a particular shift Δ. By varying the shift Δ, we can control the coherence between ψ_i and φ_i. In Figure 4, we show an example of one such construction for $k = q = 5$. Since $\sigma \in (0,1]^k$, the worst-case pair of subspaces with overlap equal to q is obtained when we pair q identical atoms with $k-q$ orthogonal atoms. In this case, the cross-spectra attains its maximum over its entire support and equals zero otherwise. For such unions, the overlap q equals the dimension of the intersection between the subspaces. We will refer to this class of block-sparse signals as *orthoblock sparse signals*.

5.1.2 COEFFICIENT SYNTHESIS

To synthesize a point that lives in the span of the sub-dictionary $\Psi \in \mathbb{R}^{n\times k}$, we combine the elements $\{\psi_1, \ldots, \psi_k\}$ and subspace coefficients $\{\alpha(1), \ldots, \alpha(k)\}$ linearly to form

$$y_i = \sum_{j=1}^{k} \psi_j \alpha(j),$$

where $\alpha(j)$ is the subspace coefficient associated with the j^{th} column in Ψ. Without loss of generality, we will assume that the elements in Ψ are sorted such that the values of the cross-spectra are monotonically decreasing. Let $y_i^c = \sum_{j=1}^{q} \psi_j \alpha_i(j)$ be the "common component" of y_i that lies in the space spanned by the principal directions between the pair of subspaces that correspond to non-orthogonal principal angles between (Φ, Ψ) and let $y_i^d = \sum_{j=q+1}^{k} \psi_j \alpha(j)$ denote the "disjoint component" of y_i that lies in the space orthogonal to the space spanned by the first q principal directions.

For our experiments, we consider points drawn from one of the two following coefficient distributions, which we will refer to as *(M1)* and *(M2)* respectively.

5. While shift-invariant dictionaries appear in a wide range of applications of sparse recovery (Mailhè et al., 2008; Dyer et al., 2010), we introduce the idea of using shift-invariant dictionaries to create structured unions of subspaces for the first time here.

- *(M1) Uniformly Distributed on the Sphere:* Generate subspace coefficients according to a standard normal distribution and map the point to the unit sphere

$$y_i = \frac{\sum_j \psi_j \alpha(j)}{\|\sum_j \psi_j \alpha(j)\|_2}, \quad \text{where } \alpha(j) \sim \mathcal{N}(0,1).$$

- *(M2) Bounded Energy Model:* Generate subspace coefficients according to *(M1)* and rescale each coefficient in order to bound the energy in the common component

$$y_i = \frac{\tau\, y_i^c}{\|y_i^c\|_2} + \frac{(1-\tau) y_i^d}{\|y_i^d\|_2}.$$

By simply restricting the total energy that each point has in its common component, the bounded energy model *(M2)* can be used to produce ensembles with small bounding constant to test the predictions in Theorem 3.

5.2 Phase Transitions for OMP

The goal of our first experiment is to study the probability of EFS—the probability that a point in the ensemble admits exact features—as we vary both the number and distribution of points in each subspace as well as the dimension of the intersection between subspaces. For this set of experiments, we generate a union of orthoblock sparse signals, where the overlap equals the dimension of the intersection.

Along the top row of Figure 5, we display the probability of EFS for orthoblock sparse signals generated according to the coefficient model *(M1)*: the probability of EFS is computed as we vary the *overlap ratio* $\delta = q/k \in [0,1]$ in conjunction with the *oversampling ratio* $\rho = k/d \in [0,1]$, where $q = \text{rank}(\Phi_1^T \Phi_2)$ equals the dimension of the intersection between the subspaces, and d is the number of points per subspace. Along the bottom row of Figure 5, we display the probability of EFS for orthoblock sparse signals generated according to the coefficient model *(M2)*: the probability of EFS is computed as we vary the overlap ratio δ and the amount of energy $\tau \in [0,1)$ each point has within its common component. For these experiments, the subspace dimension is set to $k = 20$ (left) and $k = 50$ (right). To see the phase boundary that arises when we approach critical sampling (i.e., $\rho \approx 1$), we display our results in terms of the logarithm of the oversampling ratio. For these experiments, the results are averaged over 500 trials.

As our theory predicts, the oversampling ratio has a strong impact on the degree of overlap between subspaces that can be tolerated before EFS no longer occurs. In particular, as the number of points in each subspace increases (covering radius decreases), the probability of EFS obeys a second-order phase transition, that is, there is a graceful degradation in the probability of EFS as the dimension of the intersection increases. When the pair of subspaces are densely sampled, the phase boundary is shifted all the way to $\delta = 0.7$, where70% of the dimensions of each subspace intersect. This is due to the fact that as each subspace is sampled more densely, the covering radius becomes sufficiently small to ensure that even when the overlap between planes is high, EFS still occurs with high probability. In contrast, when the subspaces are critically sampled, that is, the number of points per subspace $d \approx k$, only a small amount of overlap can be tolerated, where $\delta < 0.1$. In addition to shifting the phase boundary, as the oversampling ratio increases, the width of the transition region (where the probability of EFS goes from zero to one) also increases.

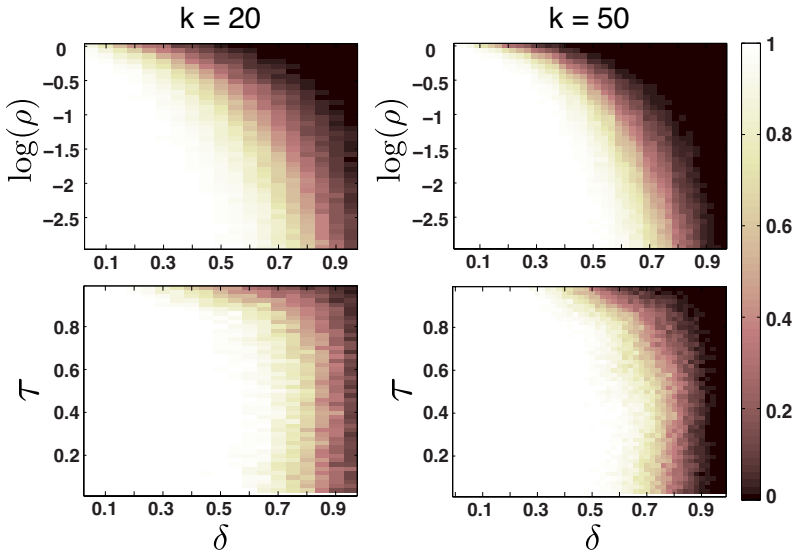

Figure 5: Probability of EFS for different coefficient distributions. The probability of EFS for a union of two subspaces of dimension $k = 20$ (left column) and $k = 50$ (right column). The probability of EFS is displayed as a function of the overlap ratio $\delta \in [0,1)$ and the logarithm of the oversampling ratio $\log(\rho)$ (top row) and the mutual energy $\tau = \|y_c\|_2$ (bottom row) .

Along the bottom row of Figure 5, we study the impact of the bounding constant on EFS, as discussed in Section 4.2. In this experiment, we fix the oversampling ratio to $\rho = 0.1$ and vary the common energy τ in conjunction with the overlap ratio δ. By reducing the bounding constant of the union, the phase boundary for the uniformly distributed data from model *(M1)* is shifted from $\delta = 0.45$ to $\delta = 0.7$ for both $k = 20$ and $k = 50$. This result confirms our predictions in the discussion of Theorem 3 that by reducing the amount of energy that points have in their subspace intersections EFS will occur for higher degrees of overlap. Another interesting finding of this experiment is that, once τ reaches a threshold, the phase boundary remains constant and further reducing the bounding constant has no impact on the phase transitions for EFS.

5.3 Comparison of OMP and NN

In this section, we compare the probability of EFS for feature selection with OMP and nearest neighbors (NN). First, we compare the performance of both feature selection methods for unions with different cross-spectra. Second, we compare the phase transitions for unions of orthoblock sparse signals as we vary the overlap and oversampling ratio.

For our experiments, we generate pairs of subspaces with structured cross-spectra as described in Section 5.1.1. The cross-spectra arising from three different unions of block-sparse signals are displayed along the top row of Figure 6. On the left, we show the cross-spectra for a union of orthoblock sparse signals with overlap ratio $\delta = 0.75$, where $q = 15$ and $k = 20$. The cross-spectra obtained by pairing shifted Lorentzian and exponential atoms are displayed in the middle and right

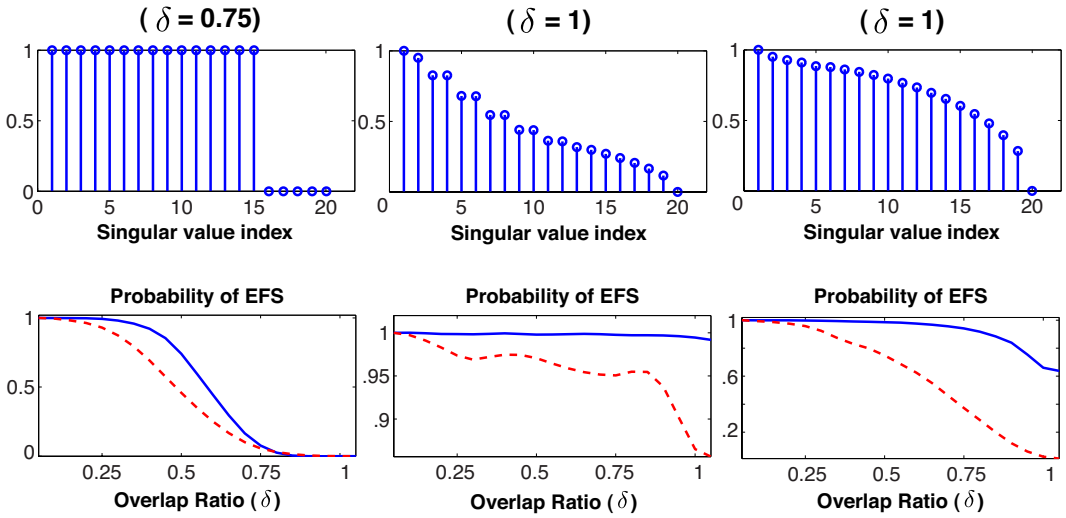

Figure 6: Probability of EFS for unions with structured cross-spectra. Along the top row, we show the cross-spectra for different unions of block-sparse signals. Along the bottom row, we show the probability of EFS as we vary the overlap ratio $\delta \in [0,1]$ for OMP (solid) and NN (dash).

columns, respectively. Along the bottom row of Figure 6, we show the probability of EFS for OMP and NN for each of these three subspace unions as we vary the overlap q. To do this, we generate subspaces by setting their cross-spectra equal to the first q entries equal to the cross-spectra in Figure 6 and setting the remaining $k-q$ entries of the cross-spectra equal to zero. Each subspace cluster is generated by sampling $d = 100$ points from each subspace according to the coefficient model *(M1)*.

This study provides a number of interesting insights into the role that higher-order principal angles between subspaces play in feature selection for both sparse recovery methods and NN. First, we observe that the gap between the probability of EFS for OMP and NN is markedly different for each of the three unions. In the first union of orthoblock sparse signals, the probability of EFS for OMP lies strictly above that obtained for the NN method, but the gap between the performance of both methods is relatively small. In the second union, both methods maintain a high probability of EFS, with OMP admitting nearly perfect feature sets even when the overlap ratio is maximal. In the third union, we observe that the gap between EFS for OMP and NN is most pronounced. In this case, the probability of EFS for NN sets decreases to 0.1, while OMP admits a very high probability of EFS, even when the overlap ratio is maximal. in summary, we observe that when data is distributed uniformly with respect to all of the principal directions between a pair of subspaces and the cross-spectra is sub-linear, then EFS may be guaranteed with high probability for all points in the set provided the sampling density is sufficiently high. This is in agreement with the discussion of EFS bounded unions in Section 4.2. Moreover, these results further support our claims that in order to truly understand and predict the behavior of endogenous sparse recovery from unions of subspaces, we require a description that relies on the entire cross-spectra.

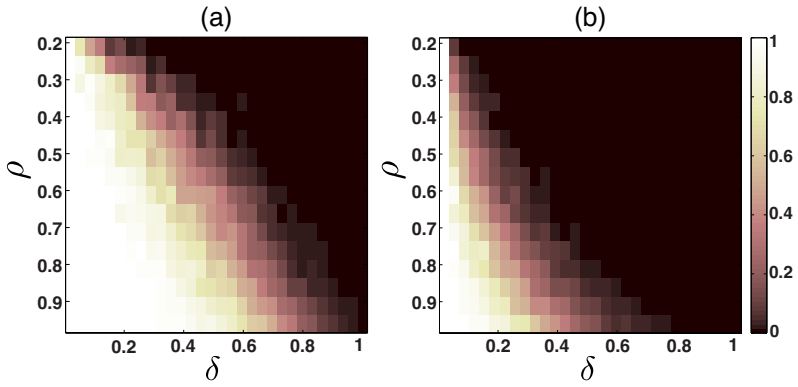

Figure 7: Phase transitions for OMP and NN. The probability of EFS for orthoblock sparse signals for OMP (a) and NN (b) feature sets as a function of the oversampling ratio $\rho = k/d$ and the overlap ratio $\delta = q/k$, where $k = 20$.

In Figure 7, we display the probability of EFS for OMP (left) and sets of NN (right) as we vary the overlap and the oversampling ratio. For this experiment, we consider unions of orthoblock sparse signals living on subspaces of dimension $k = 50$ and vary $\rho \in [0.2, 0.96]$ and $\delta \in [1/k, 1]$. An interesting result of this study is that there are regimes where the probability of EFS equals zero for NN but occurs for OMP with a non-trivial probability. In particular, we observe that when the sampling of each subspace is sparse (the oversampling ratio is low), the gap between OMP and NN increases and OMP significantly outperforms NN in terms of their probability of EFS. Our study of EFS for structured cross-spectra suggests that the gap between NN and OMP should be even more pronounced for cross-spectra with superlinear decay.

5.4 Clustering Illumination Subspaces

In this section, we compare the performance of sparse recovery methods, that is, BP and OMP, with NN for clustering unions of *illumination subspaces* arising from a collection of images of faces under different lighting conditions. By fixing the camera center and position of the persons face and capturing multiple images under different lighting conditions, the resulting images can be well-approximated by a 5-dimensional subspace (Ramamoorthi, 2002).

In Figure 2, we show three examples of the subspace affinity matrices obtained with NN, BP, and OMP for two different faces under 64 different illumination conditions from the Yale Database B (Georghiades et al., 2001), where each image has been subsampled to 48×42 pixels, with $n = 2016$. In all of the examples, the data is sorted such that the images for each face are placed in a contiguous block.

To generate the NN affinity matrices in the left column of Figure 2, we compute the absolute normalized inner products between all points in the data set and then threshold each row to select the $k = 5$ nearest neighbors to each point. To generate the OMP affinity matrices in the right column, we employ Step 1 of Algorithm 2 with the maximum sparsity set to $k = 5$. To generate the BP affinity matrices in the middle column, we solved the BP denoising (BPDN) problem in (3) via a homotopy algorithm where we vary the noise parameter κ and choose the smallest value of κ that produces

		Full-data			Half-data			Quarter-data		
		OMP	ℓ_1	NN	OMP	ℓ_1	NN	OMP	ℓ_1	NN
EFS	Mean	55.48	**73.48**	60.71	39.74	**52.50**	39.50	21.27	**28.78**	14.89
(%)	Median	55.91	**75.00**	62.5	39.06	**53.13**	39.68	18.36	**28.13**	12.50
Clustering error	Mean	**1.43**	3.69	22.03	**4.91**	13.65	37.42	**10.18**	20.14	39.63
(%)	Median	**0.78**	**0.78**	15.63	**3.13**	6.25	42.19	**6.25**	18.75	40.63

Table 1: Classification and EFS rates for illumination subspaces. Shown are the aggregate results obtained over $\binom{38}{2}$ pairs of subspaces.

$k \leq 5$ coefficients.[6] The resulting coefficient vectors are then stacked into the rows of a matrix C and the final subspace affinity W is computed by symmetrizing the coefficient matrix, $W = |C| + |C^T|$.

After computing the subspace affinity matrix for each of these three feature selection methods, we employ a spectral clustering approach which partitions the data based upon the eigenvector corresponding to the smallest nonzero eigenvalue of the graph Laplacian of the affinity matrix (Shi and Malik, 2000; Ng et al., 2002). For all three feature selection methods, we obtain the best clustering performance when we cluster the data based upon the graph Laplacian instead of the normalized graph Laplacian (Shi and Malik, 2000). In Table 1, we display the percentage of points that resulted in EFS and the classification error for all pairs of $\binom{38}{2}$ subspaces in the Yale B database. Along the top row, we display the mean and median percentage of points that resulted in EFS for the full data set (all 64 illumination conditions), half of the data set (32 illumination conditions selected at random in each trial), and a quarter of the data set (16 illumination conditions selected at random in each trial). Along the bottom row, we display the clustering error (percentage of points that were incorrectly classified) for SSC-OMP, SSC, and NN-based clustering (spectral clustering of the NN affinity matrix).

While both sparse recovery methods (BPDN and OMP) admit EFS rates that are comparable to NN on the full data set, we find that sparse recovery methods provide higher rates of EFS than NN when the sampling of each subspace is sparse, that is, the half and quarter data sets. These results are also in agreement with our experiments on synthetic data. A surprising result is that SSC-OMP provides better clustering performance than SSC on this particular data set, even though BP provides higher rates of EFS.

6. Discussion

In this section, we provide insight into the implications of our results for different applications of sparse recovery and compressive sensing. Following this, we end with some open questions and directions for future research.

6.1 "Data Driven" Sparse Approximation

The standard paradigm in signal processing and approximation theory is to compute a representation of a signal in a fixed and pre-specified basis or overcomplete dictionary. In most cases, the dictio-

6. We also studied another variant of BPDN where we solve OMP for $k = 5$, compute the error of the resulting approximation, and then use this error as the noise parameter κ. However, this variant provided worse results than those reported in Table 1.

naries used to form these representations are designed according to some mathematical desiderata. A more recent approach has been to learn a dictionary from a collection of data, such that the data admit a sparse representation with respect to the learned dictionary (Olshausen and Field, 1997; Aharon et al., 2006).

The applicability and utility of endogenous sparse recovery in subspace learning draws into question whether we can use endogenous sparse recovery for other tasks, including approximation and compression. The question that naturally arises is, "do we design a dictionary, learn a dictionary, or use the data as a dictionary?" Understanding the advantages and tradeoffs between each of these approaches is an interesting and open question.

6.2 Learning Block-Sparse Signal Models

Block-sparse signals and other structured sparse signals have received a great deal of attention over the past few years, especially in the context of compressive sensing from structured unions of subspaces (Lu and Do, 2008; Blumensath and Davies, 2009) and in model-based compressive sensing (Baraniuk et al., 2010). In all of these settings, the fact that a class or collection of signals admit structured support patterns is leveraged in order to obtain improved recovery of sparse signals in noise and in the presence of undersampling.

To exploit such structure in sparse signals—especially in situations where the structure of signals or blocks of active atoms may be changing across different instances in time, space, etc.—the underlying subspaces that the signals occupy must be learned directly from the data. The methods that we have described for learning union of subspaces from ensembles of data can be used in the context of learning block sparse and other structured sparse signal models. The application of subspace clustering methods for this purpose is an interesting direction for future research.

6.3 Beyond Coherence

While the maximum and cumulative coherence provide measures of the uniqueness of sub-dictionaries that are necessary to guarantee exact signal recovery for sparse recovery methods (Tropp, 2004) , our current study suggests that examining the principal angles formed from pairs of sub-dictionaries could provide an even richer description of the geometric properties of a dictionary. Thus, a study of the principal angles formed by different subsets of atoms from a dictionary might provide new insights into the performance of sparse recovery methods with coherent dictionaries and for compressive sensing from structured matrices. In addition, our empirical results in Section 5.3 suggest that there might exist an intrinsic difference between sparse recovery from dictionaries that exhibit sublinear versus superlinear decay in their principal angles or cross-spectra. It would be interesting to explore whether these two "classes" of dictionaries exhibit different phase transitions for sparse recovery.

6.4 Discriminative Dictionary Learning

While dictionary learning was originally proposed for learning dictionaries that admit sparse representations of a collection of signals (Olshausen and Field, 1997; Aharon et al., 2006), dictionary learning has recently been employed for classification. To use learned dictionaries for classification, a dictionary is learned for each class of training signals and then a sparse representation of a test signal is formed with respect to each of the learned dictionaries. The idea is that the test signal will

admit a more compact representation with respect to the dictionary that was learned from the class of signals that the test signal belongs to.

Instead of learning these dictionaries independently of one another, *discriminative dictionary learning* (Mairal et al., 2008; Ramirez et al., 2010), aims to learn a collection of dictionaries $\{\Phi_1, \Phi_2, \ldots, \Phi_p\}$ that are incoherent from one another. This is accomplished by minimizing either the spectral (Mairal et al., 2008) or Frobenius norm (Ramirez et al., 2010) of the matrix product $\Phi_i^T \Phi_j$ between pairs of dictionaries. This same approach may also be used to learn sensing matrices for CS that are incoherent with a learned dictionary (Mailhè et al., 2012).

There are a number of interesting connections between discriminative dictionary learning and our current study of EFS from collections of unions of subspaces. In particular, our study provides new insights into the role that the principal angles between two dictionaries tell us about our ability to separate classes of data based upon their sparse representations. Our study of EFS from unions with structured cross-spectra suggests that the decay of the cross-spectra between different data classes provides a powerful predictor of the performance of sparse recovery methods from data living on a union of low-dimensional subspaces. This suggests that in the context of discriminative dictionary learning, it might be more advantageous to reduce the ℓ_1-norm of the cross-spectra rather than simply minimizing the maximum coherence and/or Frobenius norm between points in different subspaces. To do this, each class of data must first be embedded within a subspace, a ONB is formed for each subspace, and then the ℓ_1- norm of the cross-spectra must be minimized. An interesting question is how one might impose such a constraint in discriminative dictionary learning methods.

6.5 Open Questions and Future Work

While EFS provides a natural measure of how well a feature selection algorithm will perform for the task of subspace clustering, our empirical results suggest that EFS does not necessarily predict the performance of spectral clustering methods when applied to the resulting subspace affinity matrices. In particular, we find that while OMP obtains lower rates of EFS than BPDN on real-world data, OMP yields better clustering results on the same data set. Understanding where this difference in performance might arise from is an interesting direction for future research.

Another interesting finding of our empirical study is that the gap between the rates of EFS for sparse recovery methods and NN depends on the sampling density of each subspace. In particular, we found that for dense samplings of each subspace, the performance of NN is comparable to sparse recovery methods; however, when each subspace is more sparsely sampled, sparse recovery methods provide significant gains over NN methods. This result suggests that endogenous sparse recovery provides a powerful strategy for clustering when the sampling of subspace clusters is sparse. Analyzing the gap between sparse recovery methods and NN methods for feature selection is an interesting direction for future research.

7. Proofs

In this section, we provide proofs for main theorems in the paper.

7.1 Proof of Theorem 1

Our goal is to prove that, if (7) holds, then it is sufficient to guarantee that EFS occurs for every point in $\mathcal{Y}_k$ when OMP is used for feature selection. We will prove this by induction.

Consider the greedy selection step in OMP (see Algorithm 1) for a point y_i which belongs to the subspace cluster $\mathcal{Y}_k$. Recall that at the m^{th} step of OMP, the point that is maximally correlated with the signal residual will be selected to be included in the feature set Λ. The normalized residual at the m^{th} step is computed as

$$s^m = \frac{(I - P_\Lambda) y_i}{\|(I - P_\Lambda) y_i\|_2},$$

where $P_\Lambda = Y_\Lambda Y_\Lambda^\dagger \in \mathbb{R}^{n \times n}$ is a projector onto the subspace spanned by the points in the current feature set Λ, where $|\Lambda| = m - 1$.

To guarantee that we select a point from $\mathcal{S}_k$, we require that the following greedy selection criterion holds:

$$\max_{v \in \mathcal{Y}_k} |\langle s^m, v \rangle| > \max_{v \notin \mathcal{Y}_k} |\langle s^m, v \rangle|.$$

We will prove that this selection criterion holds at each step of OMP by developing an upper bound on the RHS (the maximum inner product between the residual and a point outside of $\mathcal{Y}_k$) and a lower bound on the LHS (the minimum inner product between the residual and a point in $\mathcal{Y}_k$).

First, we will develop the upper bound on the RHS. In the first iteration, the residual is set to the signal of interest (y_i). In this case, we can bound the RHS by the mutual coherence $\mu_c = \max_{i \neq j} \mu_c(\mathcal{Y}_i, \mathcal{Y}_j)$ across all other sets

$$\max_{y_j \notin \mathcal{Y}_k} |\langle y_i, y_j \rangle| \leq \mu_c.$$

Now assume that at the m^{th} iteration we have selected points from the correct subspace cluster. This implies that our signal residual still lies within the span of $\mathcal{Y}_k$, and thus we can write the residual $s^m = z + e$, where z is the closest point to s^m in $\mathcal{Y}_k$ and e is the remaining portion of the residual which also lies in $\mathcal{S}_k$. Thus, we can bound the RHS as follows

$$\begin{aligned}
\max_{y_j \notin \mathcal{Y}_k} |\langle s^m, y_j \rangle| &= \max_{y_j \notin \mathcal{Y}_k} |\langle z + e, y_j \rangle| \\
&\leq \max_{y_j \notin \mathcal{Y}_k} |\langle z, y_j \rangle| + |\langle e, y_j \rangle| \\
&\leq \mu_c + \max_{y_j \notin \mathcal{Y}_k} |\langle e, y_j \rangle| \\
&\leq \mu_c + \cos(\theta_0) \|e\|_2 \|y_i\|_2,
\end{aligned}$$

where θ_0 is the minimum principal angle between $\mathcal{S}_k$ and all other subspaces in the ensemble.

Using the fact that $\text{cover}(\mathcal{Y}_k) = \varepsilon/2$, we can bound the ℓ_2-norm of the vector e as

$$\begin{aligned}
\|e\|_2 &= \|s - z\|_2 \\
&= \sqrt{\|s\|_2^2 + \|z\|_2^2 - 2|\langle s, z \rangle|} \\
&\leq \sqrt{2 - 2\sqrt{1 - (\varepsilon/2)^2}} \\
&= \sqrt{2 - \sqrt{4 - \varepsilon^2}}.
\end{aligned}$$

Plugging this quantity into our expression for the RHS, we arrive at the following upper bound

$$\max_{y_j \notin \mathcal{Y}_k} |\langle s^m, y_j \rangle| \leq \mu_c + \cos(\theta_0) \sqrt{2 - \sqrt{4 - \varepsilon^2}} < \mu_c + \cos(\theta_0) \frac{\varepsilon}{\sqrt[4]{12}},$$

where the final simplification comes from invoking the following Lemma.

Lemma 1 For $0 \leq x \leq 1$,

$$\sqrt{2-\sqrt{4-x^2}} \leq \frac{x}{\sqrt[4]{12}}.$$

Proof of Lemma 1: We wish to develop an upper bound on the function

$$f(x) = 2-\sqrt{4-x^2}, \quad \text{for } 0 \leq x \leq 1.$$

Thus our goal is to identify a function $g(x)$, where $f'(x) \leq g'(x)$ for $0 \leq x \leq 1$, and $g(0) = f(0)$. The derivative of $f(x)$ can be upper bounded easily as follows

$$f'(x) = \frac{x}{\sqrt{4-x^2}} \leq \frac{x}{\sqrt{3}}, \quad \text{for } 0 \leq x \leq 1.$$

Thus, $g'(x) = x/\sqrt{3}$, and $g(x) = x^2/\sqrt{12}$; this ensures that $f'(x) \leq g'(x)$ for $0 \leq x \leq 1$, and $g(0) = f(0)$. By the Fundamental Theorem of Integral Calculus, $g(x)$ provides an upper bound for $f(x)$ over the domain of interest where, $0 \leq x \leq 1$. To obtain the final result, take the square root of both sides, $\sqrt{2-\sqrt{4-x^2}} \leq \sqrt{x^2/\sqrt{12}} = x/\sqrt[4]{12}$. □

Second, we will develop the lower bound on the LHS of the greedy selection criterion. To ensure that we select a point from $\mathcal{Y}_k$ at the first iteration, we require that y_i's nearest neighbor belongs to the same subspace cluster. Let y^i_{nn} denote the nearest neighbor to y_i

$$y^i_{nn} = \arg\max_{j \neq i} |\langle y_i, y_j \rangle|.$$

If y^i_{nn} and y_i both lie in $\mathcal{Y}_k$, then the first point selected via OMP will result in EFS.

Let us assume that the points in $\mathcal{Y}_k$ admit an ε-covering of the subspace cluster $\mathcal{S}_k$, or that $\text{cover}(\mathcal{Y}_k) = \varepsilon/2$. In this case, we have the following bound in effect

$$\max_{y_j \in \mathcal{Y}_k} |\langle s^m, y_j \rangle| \geq \sqrt{1-\frac{\varepsilon^2}{4}}.$$

Putting our upper and lower bound together and rearranging terms, we arrive at our final condition on the mutual coherence

$$\mu_c < \sqrt{1-\frac{\varepsilon^2}{4}} - \cos(\theta_0)\frac{\varepsilon}{\sqrt[4]{12}}.$$

Since we have shown that this condition is sufficient to guarantee EFS at each step of Algorithm 1 provided the residual stays in the correct subspace, Theorem 1 follows by induction. □

7.2 Proof of Theorem 3

To prove Theorem 3, we will assume that the union of subspaces is bounded in accordance with (11). This assumption enables us to develop a tighter upper bound on the mutual coherence between any residual signal $s \in \mathcal{S}_i$ and the points in $\mathcal{Y}_j$. Since $s \in \mathcal{S}_i$, the residual can be expressed as $s = \Phi_i \alpha$, where $\Phi_i \in \mathbb{R}^{n \times k_i}$ is an ONB that spans $\mathcal{S}_i$ and $\alpha = \Phi_i^T s$. Similarly, we can write each point in $\mathcal{Y}_j$ as

$y = \Phi_j \beta$, where $\Phi_j \in \mathbb{R}^{n \times k_j}$ is an ONB that spans $\mathcal{S}_j$, $\beta = \Phi_j^T y$. Let $\mathcal{B}_j = \{\Phi_j^T y_i\}_{i=1}^{d_j}$ denote the set of all subspace coefficients for all $y_i \in \mathcal{Y}_j$.

The coherence between the residual and a point in a different subspace can be expanded as follows:

$$\begin{aligned}
\max_{y \in \mathcal{Y}_j} \frac{|\langle s, y \rangle|}{\|s\|_2} &= \max_{\beta \in \mathcal{B}_j} \frac{|\langle \Phi_i \alpha, \Phi_j \beta \rangle|}{\|\alpha\|_2} \\
&= \max_{\beta \in \mathcal{B}_j} \frac{|\langle \alpha, \Phi_i^T \Phi_j \beta \rangle|}{\|\alpha\|_2} \\
&= \max_{\beta \in \mathcal{B}_j} \frac{|\langle \alpha, U \Sigma V^T \beta \rangle|}{\|\alpha\|_2} \\
&= \max_{\beta \in \mathcal{B}_j} \frac{|\langle U^T \alpha, \Sigma V^T \beta \rangle|}{\|\alpha\|_2} \\
&\leq \max_{\beta \in \mathcal{B}_j} \frac{\|U^T \alpha\|_\infty}{\|\alpha\|_2} \|\Sigma V^T \beta\|_1, \qquad (13)
\end{aligned}$$

where the last step comes from an application of Holder's inequality, that is, $|\langle w, z \rangle| < \|w\|_\infty \|z\|_1$.

Now, we tackle the final term in (13), which we can write as

$$\max_{\beta \in \mathcal{B}_j} \|\Sigma V^T \beta\|_1 = \max_{y \in \mathcal{Y}_j} \|\Sigma V^T \Phi_j^T y\|_1 = \max_{y \in \mathcal{Y}_j} \|\Sigma (\Phi_j V)^T y\|_1,$$

where the matrix $\Phi_j V$ contains the principal vectors in subspace $\mathcal{S}_j$. Thus, this term is simply a sum of weighted inner products between the principal vectors $\Phi_j V$ and all of the points in $\mathcal{S}_j$, where Σ contains the cross-spectra in its diagonal entries.

Since we have assumed that the union is bounded, this implies that the inner product between the first q principal vectors and the points in $\mathcal{Y}_j$ are bounded by γ, where $q = \|\sigma_{ij}\|_0 = \operatorname{rank}(G)$. Let $\Phi_j V_q \in \mathbb{R}^{n \times q}$ be the first q singular vectors of G corresponding to the nonzero singular values in Σ and let $\Sigma_q \in \mathbb{R}^{q \times q}$ be a diagonal matrix with the first q nonzero singular values of G along its diagonal. It follows that $\|\Sigma (\Phi_j V)^T y\|_\infty = \|\Sigma_q (\Phi_j V_q)^T y\|_\infty \leq \gamma$. Now, suppose that the bounding constant $\gamma < \sqrt{1/q}$. In this case,

$$\max_{y \in \mathcal{Y}_j} \|\Sigma (\Phi_j V)^T y\|_1 \leq \gamma \|\sigma_{ij}\|_1.$$

Note that for bounded unions of subspaces, the term on the right can be made small by requiring that the bounding constant $\gamma \ll 1$. Plugging this bound into (13), we obtain the following expression

$$\max_{y \in \mathcal{Y}_j} \frac{|\langle s, y \rangle|}{\|r\|_2} \leq \gamma \|\sigma_{ij}\|_1 \frac{\|U^T \alpha\|_\infty}{\|\alpha\|_2} = \gamma \|\sigma_{ij}\|_1 \|U\|_{2,2} = \gamma \|\sigma_{ij}\|_1,$$

where this last simplification comes from the fact that U is unitary and has spectral norm equal to one. Note that this bound on the mutual coherence is informative only when $\gamma \|\sigma_{ij}\|_1 < \sigma_{\max} \leq 1$. This completes the proof. □

Acknowledgments

Thanks to Dr. Chinmay Hegde, Dr. Ankit Patel, Mahdi Soltanolkotabi, and Dr. Christoph Studer for helpful discussions and comments on this paper. Thanks also to Dr. Arian Maleki and Dr. Joel Tropp for helpful discussions. We would like to thank the anonymous reviewers, whose comments and suggestions were invaluable. ED was supported by a NSF GRFP 0940902 and a Texas Instruments Distinguished Graduate Fellowship. ACS and RGB were partially supported by following grants: NSF CCF-1117939, CCF-0431150, CCF-0728867, CCF-0926127; DARPA N66001-11-1-4090, N66001-11-C-4092; ONR N00014-08-1-1112, N00014-10-1-0989; AFOSR FA9550-09-1-0432; ARO MURIs W911NF-07-1-0185 and W911NF-09-1-0383.

References

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54(11):4311–4322, 2006.

E. Arias-Castro, G. Chen, and G. Lerman. Spectral clustering based on local linear approximations. *Electron. J. Stat.*, 5(0):217–240, 2011.

R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, 56(4):1982–2001, 2010.

R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(2):218–233, February 2003.

T. Blumensath and M. Davies. Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Trans. Inform. Theory*, 55(4):1872–1882, 2009.

G. Chen and G. Lerman. Spectral curvature clustering. *Int. J. Computer Vision*, 81:317–330, 2009.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comp.*, 20(1):33–61, 1998.

G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *SPIE J. Opt. Engin.*, 33(7):2183–2191, 1994.

E. L. Dyer. Endogenous sparse recovery. Master's thesis, Electrical & Computer Eng. Dept., Rice University, October 2011.

E. L. Dyer, M. Duarte, D. J. Johnson, and R. G. Baraniuk. Recovering spikes from noisy neuronal calcium signals via structured sparse approximation. *Proc. Int. Conf. on Latent Variable Analysis and Sig. Separation*, pages 604–611, 2010.

E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, June 2009.

E. Elhamifar and R. Vidal. Clustering disjoint subspaces via sparse representation. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, pages 1926–1929, March 2010.

E. Elhamifar and R. Vidal. Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Machine Intell.*, 2013.

A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(6):643–660, 2001.

B. V. Gowreesunker, A. Tewfik, V. Tadipatri, J. Ashe, G. Pellize, and R. Gupta. A subspace approach to learning recurrent features from brain activity. *IEEE Trans. Neur. Sys. Reh*, 19(3):240–248, 2011.

K. Kanatani. Motion segmentation by subspace separation and model selection. In *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, 2001.

Y. Lu and M. Do. Sampling signals from a union of subspaces. *IEEE Sig. Proc. Mag.*, 25(2):41–47, March 2008.

B. Mailhè, S. Lesage, R. Gribonval, F. Bimbot, and P. Vandergheynst. Shift-invariant dictionary learning for sparse representations: extending K-SVD. In *Proc. Europ. Sig. Processing Conf. (EUSIPCO)*, 2008.

B. Mailhè, D. Barchiesi, and M. D. Plumbley. INK-SVD: Learning incoherent dictionaries for sparse representations. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, pages 3573 –3576, March 2012.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Proc. IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, June 2008.

A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Proc. Adv. in Neural Processing Systems (NIPS)*, 2:849–856, 2002.

B. Olshausen and D. Field. Sparse coding with an overcomplete basis set: a strategy employed by V1. *Vision Res.*, 37:3311–3325, 1997.

R. Ramamoorthi. Analytic PCA construction for theoretical analysis of lighting variability in images of a lambertian object. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(10):1322–1333, 2002.

I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Proc. IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, pages 3501–3508, June 2010.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(8):888–905, August 2000.

M. Soltanolkotabi and E. J. Candès. A geometric analysis of subspace clustering with outliers. *Annals of Statistics*, 40(4):2195–2238, 2012.

M. Soltanolkotabi, E. Elhamifar, and E. J. Candès. Robust subspace clustering. *CoRR*, abs/1301.2603, 2013.

J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 52(3):1030 –1051, March 2006.

R. Vidal. Subspace clustering. *IEEE Sig. Proc. Mag.*, 28(2):52–68, 2011.

R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Trans. Pattern Anal. Machine Intell.*, 27(12):1945–1959, 2005.

Y. Wang and H. Xu. Noisy sparse subspace clustering. *Proc. Int. Conf. Machine Learning*, 2013.

J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Proc. European Conf. Comp. Vision (ECCV)*, 2006.

T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. *Int. J. Computer Vision*, 100(3):217–240, 2012.